MCI

MCI Telecommunications Corporation
1801 Pennsylvania Avenue, NW
403E
Washington, DC 20006

Lisa R. Youngers
Regulatory Attorney

January 27, 1999

Magalie Roman Salas
Secretary
Federal Communications Commission
1919 M Street, N.W. Room 222
Washington, D.C. 20554

Re: In the Matter of Performance Measurements and Reporting Requirements for Operations Support Systems, Interconnection, and Operator Services and Directory Assistance
CC Docket No. 98-56, RM-9101
Ex Parte Filing

Dear Ms. Salas:

Enclosed is a copy of MCI WorldCom's Post-Workshop Brief on Statistical Testing filed in the Louisiana Public Service Commission's Docket No.U-22252, Subdocket C In Re: BellSouth Telecommunications, Inc. Service Quality Performance Measurements. Also enclosed is a copy of "Using Permutation Tests to Evaluate the Significance of CLEC vs. ILEC Service Quality Differentials" by John D. Jackson of the E Group, Economics and Econometrics Research Consultants. This paper was also filed in the same Louisiana Service Quality docket mentioned above, included as Attachment 6 in the November 19, 1998 joint filing by AT&T Communications of the South Central States, MCI Telecommunications Corporation, and Sprint Communications Company L.P. Please place both of these documents in the above-referenced FCC docket.

Pursuant to Section 1.1206(b) of the Commission's Rules, two copies of this brief and paper are being submitted to the Secretary.

Sincerely,

Lisa R Youngers

Lisa R. Youngers
Regulatory Counsel

cc: Daniel R. Shiman,
Common Carrier Bureau

BEFORE THE

LOUISIANA PUBLIC SERVICE COMMISSION

LOUISIANA PUBLIC SERVICE COMMISSION, EX PARTE	DOCKET NO. U-22252 SUBDOCKET C

IN RE: BELLSOUTH
TELECOMMUNICATIONS, INC., SERVICE
QUALITY PERFORMANCE
MEASUREMENTS

MCI TELECOMMUNICATIONS CORPORATION'S POST-WORKSHOP BRIEF ON STATISTICAL TESTING

MCI Telecommunications Corporation ("MCI") hereby submits the following post-workshop brief relating to statistical testing. This brief is being submitted pursuant to the Louisiana Public Service Commission's ("LPSC") Notice, dated December 4, 1998. As set forth below, MCI urges the LPSC staff to reject the statistical test proposed by BellSouth Telecommunications, Inc. ("BellSouth"). BellSouth's proposed test does not consider the necessary aspects of parity, can be used to cover up actual parity violations, and relies on a "suite of tests" that fail to yield a swift and objective parity determination.

I. INTRODUCTION

The LPSC issued a General Order, dated August 31, 1998, adopting service quality performance measurements ("SQM") to be reported by BellSouth Telecommunications, Inc. ("BellSouth"). The LPSC's General Order also directed the staff to conduct a series of workshops to further investigate several SQM issues.

The LPSC staff conducted a workshop on November 30 and December 1, 1998 to address the various statistical testing methodologies proposed by the parties. Under the General Order, BellSouth was to perform the statistical test that it proposed (which at that time was statistical process control),[1] the modified z-test endorsed by the Local Competition Users Group ("LCUG"), and the pooled variance test offered by the Federal Communications Commission ("FCC") in its Notice of Proposed Rulemaking. The LPSC directed BellSouth to perform this statistical testing so that the competence of each test could be evaluated by the LPSC staff and interested parties at the workshop. Unfortunately, BellSouth's statistical analysis was primarily a one-sided presentation of its new methodology.[2] Thus, there was insufficient results from the modified z-test and FCC pooled variance test for evaluation and comparison to BellSouth's proposed methodology. MCI encourages BellSouth to provide sufficient results in its next report so that the staff and the interested parties may make a

[1] BellSouth abandoned statistical process control as its proposed statistical testing methodology in Louisiana and set forth an alternative testing methodology at the workshop.

[2] BellSouth claimed that it did not have the necessary information to perform the z-test and pooled variance test on all of the measures presented. For example, for the Average OSS Response Interval Measurement, BellSouth used daily summary averages to perform its proposed statistical test. But, the z-test and pooled variance test supposedly could not be performed with such data. Even assuming BellSouth's claim was true, MCI believes the necessary information to perform the z-test and pooled variance test was available to BellSouth. Surely, the underlying data from which the daily summary averages were calculated was available. Further testing using the z-test and pooled variance test will be presented by BellSouth in February as it is to perform statistical testing on an additional five measures. <u>See</u> LPSC Notice, dated December 4, 1998, p. 5. Hopefully, BellSouth's next report will provide sufficient information for the LPSC staff to compare the testing methodologies.

meaningful comparison of the methodologies. Additionally, BellSouth's providing of the raw data to the CLECs for analysis would provide these parties with the ability to compare the results of the various tests.

At the workshop, BellSouth presented its alternative testing methodology developed by Ernst & Young ("EY") statisticians, Fritz Scheuran, Susan Hinkins, and Ed Mulrow. For a particular SQM, the main part of the test involves:

(1) disaggregating across several attributes of the SQM (such as dispatch vs. non-dispatch, business vs. residential, number of circuits, new order vs. change order, first vs. second half of the month, and individual wire centers);

(2) weighting of the ILEC observations in each category to make them similar to the relevant CLEC observations;

(3) computing the weighted ILEC and the CLEC means differences at the individual wire center level by summing the adjusted means differences across the categories above;

(4) randomly assigning each wire center's means difference to one of thirty groups to remove geographic effects;

(5) computing the average of the means differences for each group, using these thirty averages to compute the variance about the weighted means difference;

(6) dividing the variance by the square root of 30, and then dividing such figure into the weighted means difference; and

(7) comparing the resulting statistic to a t distribution with 29 degrees of freedom. The full procedure involves using the main part of the test as described above in conjunction with several graphical analyses to provide an indication of whether the provision of the service might be out of parity.[3]

Dr. Colin Mallows, AT&T's statistical expert, had prepared a list of eight criticisms of the BellSouth approach:

(1) The conclusions of the BellSouth team are not supported by the evidence that they have presented;

(2) BellSouth did not provide suitable data to perform the required tests on Average OSS Response Interval;

(3) BellSouth misstates the power of its test;

(4) The conclusion of the BellSouth analysts that the FCC and LCUG procedures have to rely on unwarranted assumptions is incorrect;

(5) The BellSouth analysts use the wrong variance estimator in its Replicate Variance Estimation, leading to a variance estimate that is perhaps double what it should be;

(6) BellSouth ignores within wire center variation;

[3] The utter subjectiveness of BellSouth's approach is discussed in further detail below. At one point in its presentation, BellSouth goes so far as to recommend the use of customer surveys to determine whether "statistically significant" differences are even important. See BellSouth Interim Statistical Analysis, p. 7. MCI contends that such subjective analysis are inappropriate and could lead to uncertainty and confusion.

(7) The t statistic upon which BellSouth bases its test is not a test of the hypothesis that the BellSouth and CLEC populations are the same; and

(8) BellSouth's criticism of the LCUG multiple testing procedure likely overstates the degree of interdependence among the SQM tests.

MCI is in complete agreement with each of Dr. Mallows' criticisms. Indeed, the EY consultants also agreed with several of the criticisms and appear willing to revise the BellSouth proposal.

In addition to the criticisms of Dr. Mallows, MCI sets forth below, additional, fundamental problems with BellSouth's approach. These flaws in BellSouth's methodology cannot be corrected with the minor adjustments that the EY consultants agreed to consider. Rather, it is the core methodology of the test that makes the proposal inappropriate for measuring whether true parity exists. Due to these fundamental problems, MCI urges the LPSC staff to reject BellSouth's proposed statistical methodology. The modified z-test continues to be the most appropriate test to determine whether BellSouth is providing parity service to CLECs and should be adopted by this Commission.

II. ARGUMENT

MCI's objections to the BellSouth approach to parity testing can be classified into three main categories: (1) Problems with the BellSouth procedure -- these include problems with weighting, aggregation, needless splitting of categories, and BellSouth's proposed "suite of tests"; (2) Problems with the BellSouth test, itself -- these include testing for means differences vs. testing for parity, ignoring within wire center variation, using the wrong measure of between group variance, and insistence on two-tailed testing; and (3) Problems with the BellSouth criticism of the LCUG/FCC methodology -- these include BellSouth's analysis of the independence requirement, its suggestion of inherent differences in underlying assumptions for the three approaches, and its suggestion that the artificial corrections for potentially confounding factors affecting SQMs.[4]

A. Problems With The BellSouth Procedure

1. Weighting

MCI will first consider some problems inherent in the "weighting" aspect of BellSouth's "weight and aggregate" philosophy. BellSouth suggests that attempts to analyze an essentially heterogeneous SQM without making BellSouth's suggested adjustments[5] may result in means differences that do not in fact exist. As an illustration, BellSouth offered an example in its Interim Statistical Analysis, the excerpts of which are attached as Exhibit 1.

In the attached example, there should be no difference in the Order Completion Interval means for BellSouth and the CLECs, since the mean times for new orders and for

[4] As discussed by MCI below, if such factors do indeed obfuscate the true service provision relationship, the LCUG/FCC approach can correct for them in a much less arbitrary way that does the BellSouth approach.

[5] BellSouth proposes "disaggregating" according to component attributes, weighting these heterogeneous attributes so that the ILEC sample more closely resembles the CLEC sample, reaggregating the weighted data, and analyzing the adjusted variables rather than the raw data counterparts.

change orders are the same for both -- 2 days for new orders and 1 day for change orders. However, if one computes the actual averages implied by the data in the example, there is an apparent difference where none should exist -- 1.25 for Provider A and 1.67 for Provider B. This difference is due to the volume of orders in each category -- 30 new orders and 90 change orders for Provider A and 60 new orders and 30 change orders for Provider B.

In Appendix B of the Interim Statistical Analysis (see Exhibit A for relevant excerpt), BellSouth points out that if one were to simply weight each group means difference by the percentage of CLEC observations in that group [(60/30)=2 for new orders and (30/90)=1/3) for change orders], this apparent difference vanishes.[6] BellSouth then indicates that this procedure is strictly equivalent to (1) weighting the BellSouth observations to make them more closely resemble the CLECs, and (2) computing a weighted average for BellSouth to compare to the unweighted CLEC average.

In BellSouth's example, the weights are the ratio of CLEC observations to BellSouth observations in each group, and the weighting of each observation is accomplished by multiplying the BellSouth group average by the number of BellSouth observations in that group. The divisor of the weighted average turns out to be the total number of CLEC observations for that SQM.[7] By weighting BellSouth new orders by a factor of 2, the BellSouth change orders by a factor of 1/3, and computing the implied weighted average, the

[6] D^=((60)(2-2) + (30)(1-1))/(30 + 60)=0

[7] BST Mean = (((60/30)*30*2) + ((30/90)*90*1))/30 + 60 = 150/90=1.67

BellSouth adjusted mean is converted to the CLEC mean. A direct comparison of the BellSouth mean with the CLEC mean reveals no difference in service performance -- precisely as should be the case. It may be useful to think of this type of experiment as putting ILEC observations in a CLEC world.

MCI wishes to emphasize three important points relating to this type of weighting procedure. First, the reason that weighting is necessary -- and BellSouth has not been clear about this point -- is **not** because of the technical production processes that result in new orders taking (hypothetically) two minutes to complete and change orders taking (hypothetically) only one minute. Rather, it is because the distribution of BellSouth orders across these two order categories differs from the distribution of CLEC orders.

To illustrate this point, assume that BellSouth has the same order distribution as in the previous example (i.e., 30 new orders, 90 change orders), but now assume the CLECs have the same percentage distribution (i.e., 15 new orders and 45 change orders). It still takes two minutes to complete a new order and one minute to complete a change order for both BellSouth and CLECs on average. Under these assumptions, the CLEC mean is 1.25,[8] which is exactly the same as the unweighted BellSouth average computed in BellSouth's example. Clearly then, the need to differentially weight sub-categories of SQMs hinges on whether BellSouth and CLEC observations are differentially distributed across these sub-categories, and not on sub-category differences *per se*. Of course, the total number of orders in each

[8] (15 * 2) + (45 * 1)/ (15 + 45) = 1.25

sub-category will differ for BellSouth and the CLECs, presumably in proportion to the total number of orders placed with that provider. But, there is no reason *a priori* why the percentage distribution of orders should differ for BellSouth and the CLECs across sub-categories. Thus, there is no compelling justification for BellSouth differentially weighting measures across sub-categories.

A second problem with BellSouth's weighting procedures is that its weights are not unique. That is, there are other, equally reasonable, weighting schemes that could be applied to the sub-categories that produce different means and variances than the BellSouth weights, but yield the same general result. For example, suppose the CLECs are put in an ILEC world. This could be accomplished by weighting the CLEC observations in each group by the ratio of BellSouth to CLEC observations in that group and dividing the resulting sum by the total number of BellSouth observations.[9] The resulting adjusted CLEC mean is exactly the same as the unadjusted BellSouth mean. The weights are different [3 and 1/2 in this example vs. 2 and 1/3 in BellSouth's example], the means (and variances) are different, but the end result -- that weighting has adjusted the mean of one provider to equal the unadjusted mean of the other provider -- is the same. MCI sees no reason to favor BellSouth's weighting scheme to this one.

[9] Thus, using BellSouth's initial example, the CLEC mean = ((30/60) * 60 * 2) + (90/30 * 30 * 1)) / (30 + 90) = 150/120 = 1.25

Indeed, there are any number of equally valid weighting schemes. One could weight both CLEC and BellSouth observations by the relative size of the relevant sub-category expressed as a percent of the total sample. In this case both CLEC and BellSouth adjusted means turn out to be 1.4286 [=(90*2+120*4)/210]. From an economic perspective, one could easily justify weighting each sub-category by the percent of total revenue, total cost, or total profit accruing to the particular SQM that is attributable to that subcategory. The point being that there are many weighting schemes that could be used in place of BellSouth's scheme. Each of these schemes will produce different means and different variances for the weighted data. To suggest that the BellSouth weighting scheme is somehow not arbitrary, as was done in the meeting, is misleading, at best. This arbitrary choice from a multiplicity of weighting schemes is even more important when one considers MCI's third problem with BellSouth's proposed weighting scheme.

The BellSouth weighting scheme, and any other weighting scheme for that matter, can be used to cover up actual means differences. To illustrate this point, let us change the initial BellSouth illustration slightly. Let the distribution of observations across new and change orders remain the same for BellSouth and the CLECs. Also, let BellSouth's mean time for new order completion remain at two minutes and order change completion time remain at one minute. However, let us now suppose that the CLEC's mean completion time for new orders is 2.25 minutes and that BellSouth's mean time for change order completion is now 0.5 minutes.

In this example, there is a clear violation of the parity requirement in the case of new orders. There is no violation of the parity requirement for change orders since they are provided to the CLECs at least as quickly as BellSouth provides them to itself. Note that the new raw data mean for the CLECs is the same value as their old mean — 1.67.[10] Additionally, the adjusted BellSouth mean is the same value — 1.67.[11] Direct comparison of the CLEC mean with BellSouth's mean reveals no difference in service performance, exactly as before. This result is reinforced by computing the adjusted means difference according to the BellSouth formula.[12] *Thus, all of BellSouth's weighted analysis implies that there is full parity in the provision of order completion intervals when there is a clear lack of parity in the completion interval for new orders.*

What makes the BellSouth weighting approach so pernicious is not only that it cannot detect this type of smoothing over the lack of parity, but that it actually encourages such obfuscation by requiring layer upon layer upon layer of such weightings. Recall that BellSouth suggested breaking the actual data on Order Completion Interval down into the following sub-categories: dispatch vs. non-dispatch (2 groups), business vs. residential vs. special (3 groups), less than 10 ckts vs. more than 10 ckts (2 groups), new orders vs. change

[10] CLEC mean = (60 * 2.25) + (30 * .05) / (60 + 30) = 150/90 = 1.67

[11] BellSouth mean = ((60/30 * 30 * 2) + (30/90 * 90 * 1)) / (30 + 60) = 150/90 = 1.67

[12] D^ = ((60 * (2-2.25)) + (30 * (1-0.5))) / (30 + 60) = (-15 + 15) / 90 = 0

orders vs. transfer orders (3 groups), and first vs. second half of the month as completion date (2 groups), and finally by wire center (230 groups). Thus, BellSouth requires that roughly 70 subcategories be weighted for each of 230 individual wire centers. Since each of these categories must be weighted and since the choice of weighting scheme is essentially arbitrary, it is not only possible, it is quite likely that ***many*** parity failures will be covered up by this procedure. Since any given category may be more important to a CLEC at a given time than another category, these types of undetected parity failures are unacceptable. For example, if MCI is launching a residential program at the beginning of a particular month, BellSouth could render poor residential service to MCI at that time. Yet, the proposed weighting scheme could cover up such poor service if BellSouth rendered good service to MCI in a category that was less important to MCI at that time.

One clue to parity failure cover-ups may be a reversal of means difference sign from a negative on the unweighted data to a positive on the weighted data. For example, in the Interim Statistical Analysis, the Order Completion Interval tests for the New Orleans LATA change from negative and insignificant when unweighted to positive and significant when weighted. It may be that the explanation for this perverse outcome lies in the obfuscation inherent in the BellSouth weighting process. Of course, MCI can only offer this conjecture since it has not had access to any of the BellSouth data.

It should be obvious that MCI has grave doubts about the legitimacy of BellSouth's weighting approach, and so should the LPSC. The inability to weight destroys BellSouth's *"weight and aggregate" testing philosophy. No party wants to aggregate without weighting.* BellSouth was on the mark when it correctly pointed out the hazards of aggregating across unweighted heterogeneous sub-categories to obtain an aggregate SQM.

In light of the inherent flaws in BellSouth's weighting procedure, the LPSC must reject BellSouth's weight and aggregate methodology. So, what kind of analysis can be legitimately performed? The answer is clear. Do not weight; simply test the disaggregated data. If this means conducting 70 LCUG z-tests for each of 230 wire centers and tabulating the results to see if the percent of tests failed is sufficient to reject parity, then that is what is necessary. There is no appropriate alternative to this approach.

2. Aggregation

While BellSouth certainly does not advocate aggregating across unweighted sub-categories, it does appear to favor certain types of aggregation that MCI finds unacceptable. One such aggregation strategy is its use of pooled CLEC data rather than individual CLEC data. Perhaps BellSouth simply pooled the CLECs for expository purposes, to avoid confidentiality violations, and, in fact, plans to conduct its comparisons CLEC by CLEC. If so, BellSouth should have stated as much in its Interim Statistical Analysis; it did not so state.

MCI believes that BellSouth plans to continue to use aggregate CLEC data in its analysis for two reasons: (i) BellSouth can avoid any small sample testing problems in its analysis (and conducting analyses CLEC by CLEC will inevitably lead to small sample testing problems for BellSouth's proposed test); and (ii) BellSouth's test requires large CLEC samples, just as the LCUG and FCC tests do (see, for example, p. B 10 in the Interim Statistical Analysis).[13] If BellSouth does plan to continue to use aggregate CLEC data in its analysis, MCI reiterates its objection to aggregating across CLECs or, for a given CLEC, over time. It is simply too easy to game the system, whether by discriminating against one key CLEC while treating all of the others well or by discriminating against the CLEC for some limited sub-period and supplying it at parity the remainder of the time. In either case, the aggregation would render the test statistics smaller than they should be, making rejection of parity less likely.

Anticipating BellSouth's argument, MCI believes that (possibly) detecting this sort of gaming and discrimination by periodic auditing of the BellSouth data is not a sufficient remedy. In the case of a CLEC attempting to enter a burgeoning competitive marketplace, justice delayed is no justice at all. These audits will be expensive and time consuming and should not take the place of appropriate performance measurements and standards, statistical testing and self-executing remedies. A better procedure is to employ the disaggregated data approach and to use the small sample testing procedure that MCI outlined in its paper on permutation testing.

3. Unjustified Expansion of Sub-Categories

[13] LCUG addresses this problem by using permutation tests.

Besides problems arising from weighting and aggregation, there also seems to be a question concerning the possibly needless splitting of the data on a given SQM into finer and finer categories. This is a particularly important question in view of BellSouth's arguments against "deep testing." For the Order Completion Interval SQM, BellSouth argues that, with deep testing, it would have to conduct about 70 tests for each of 230 wire centers, and undertake over 16,000 separate means difference tests on this SQM alone. Thus, BellSouth argues, deep testing of the highly disaggregated type which LCUG recommends is prohibitive because of the sheer number of analyses involved.

In the example given above, the number of separate means difference tests that must be performed hinges crucially on the assertion that the wire center is the appropriate level of geographic disaggregation. But, the LPSC adopted, in its General Order, that the appropriate level of geographic disaggregation is the MSA level. At such level, the required number of tests would fall to about 600 (about 70 tests for each of 9 MSAs),[14] the results of which could be tabulating to see if the percent of tests failed is sufficient to reject parity. Despite the LPSC's decision, the Interim Statistical Analysis was not based on the MSA level of disaggregation. Unfortunately, BellSouth provided no statistical evidence to support the wire center level of disaggregation used in the Interim Statistical Analysis. BellSouth's criticism of deep testing based on the assumptions set forth in the Interim Statistical Analysis is invalid. In fact, based on BellSouth's proposed "suite of tests," including the graphical

[14] This number assumes that all other sub-category divisions are retained.

analysis of the many plots and distributions, it is BellSouth's proposed test that is unworkable for purposes of determining parity.

At the workshop, the parties discussed how one might verify whether a particular sub-category division is empirically relevant. MCI suggested that one could pool the Order Completion Interval data for BellSouth and the CLECs and estimate a regression equation. The equation would consist of a constant term, a group dummy (say, d=0 for BellSouth observations, d=1 for CLEC observations), analogously defined dummies for the various sub-categories of Order Completion Interval characteristics, a set of 229 wire center dummies, and a set of interaction variables, created by multiplying the group dummy times all other included variables. Standard t and F tests are available to determine whether any of these potentially confounding factors significantly affect Order Completion Interval provision. For the case at hand, non-nested testing procedures of Davidson and MacKinnon (1981) or Cox (1961,1962) can be used to detect whether the above specification, one in which the set of wire center dummies is replaced by a set of 8 MSA dummies, is the appropriate specification. There are a variety of statistical methods for determining whether a particular Order Completion Interval classification is empirically relevant. Until BellSouth chooses to either share the data or provide such results, the LPSC should reject BellSouth's arguments concerning the prohibitive nature of deep testing.

4. The "Suite of Tests" Approach

Contrary to the LCUG approach, which recommends one simple test that has power against several alternatives, BellSouth recommends a "suite of tests," which consists of a statistical test accompanied by several graphical analyses. The graphical analyses entail frequency distributions and quantile plots presented at various levels of disaggregation noted above. It is not clear whether one plots first, then tests, or tests first, then plots. What is clear is that this "suite of tests" adds another unnecessary layer to the administrative hierarchy of decision-making.

Under BellSouth's proposal, if it finds a test indication of non-parity, then BellSouth consults its frequency distributions and scatter plots to verify this conclusion -- or vice-versa. If it still finds indications of non-parity, it "drills down" through successive levels of disaggregation until its plots tell the story it wants to hear. Most likely, BellSouth will find some anomaly in data collection, recording errors, or some other innocuous source of seeming non-parity. Perhaps, BellSouth might acknowledge the possibility of discrimination against the CLECs as a potential source, but there will always be that qualifier -- it could be something else. If left to its own devices, BellSouth may never find out-and-out discrimination; it has a vested interest in not doing so. But, vested interest or not, these drill downs take time. This time could be crucial to the ability of CLECs, as new entrants, to make inroads into the market, or to lose forever any potential for developing a customer base because of continued BellSouth discrimination. For this reason, LCUG has argued for swift and certain enforcement of parity violations. The BellSouth "suite of tests" is geared to foil such attempts, and therefore, MCI opposes such approach. MCI urges the LPSC staff to also reject such approach.

B. Problems With The BellSouth Test

The following section discusses, for the most part, criticisms raised by Dr. Mallows regarding the BellSouth statistical test. MCI defers to Dr. Mallows for detailed explanations of his critiques, but would like to share some brief comments regarding these criticisms.

First, the BellSouth test, as initially designed, tests for means differences, not parity. Parity requires that the CLEC and BellSouth samples follow the same distribution -- that is, the samples should have the same mean and the same variance. The LCUG z-statistic is a means difference test that has more power, than do the alternative tests, to reject parity if the CLEC variance exceeds the BellSouth variance. A detailed explanation for the basis of this conclusion may be found on page 5 of Joint Attachment 3, Testing for Parity in the Quality of Services Provided by ILECs to CLECs: A Comparison of Large Sample Procedures, filed by AT&T, MCI and Sprint on November 19, 1998.

On the other hand, the BellSouth test is a means difference test that allows for different CLEC and BellSouth variances. Since parity requires that the variances be the same, the BellSouth test cannot test for parity -- if the CLEC variance exceeds the BellSouth variance, parity is violated. To its credit, the BellSouth analysts seemed to accept this concept and have agreed to modify their procedure to reflect a sensitivity to departures from equality of variance. How successful they will be remains to be seen.

As Dr. Mallows points out, the BellSouth test also ignores the within group variance at the wire center level. This is a particularly important result since it differentiates the LCUG approach from the BellSouth test. The LCUG approach is to disaggregate to the point at which all relevant confounding factors have been eliminated and then perform the LCUG z-test. This means that, at the wire center level, the appropriate variance to use in computing the LCUG test statistic is the within wire center variance. BellSouth computes means differences at the wire center level, groups them randomly into 30 groups and computes group means, then computes a between group variance based on these group means. The BellSouth test, therefore, excludes consideration of the very variance that LCUG views as fundamental.

Dr. Mallows also points out that BellSouth uses the wrong estimator for its between group variance. BellSouth disagrees, but MCI agrees completely with Dr. Mallows' conclusion. The expected value of the group means is not the population analogue of D^, which is what the BellSouth analysts assume in their variance formula (p. B 8). Furthermore, as Dr. Mallows notes, the appropriate divisor in the BellSouth test formula is the standard error of D^, which, since D^ is a weighted average, is likely to be considerably smaller than the divisor used by BellSouth.

The LPSC should be made aware of the obvious impact of this error. When BellSouth reports its comparative findings for Order Completion Interval parity tests, the BellSouth statistics are about half the size of the LCUG and FCC statistics. Since this relationship holds whether one is referring to adjusted or unadjusted data, it is unlikely that weighting is responsible for this error. The numerator of all three test statistics is the same; thus, the BellSouth standard error is about twice the size of the LCUG and FCC standard errors.[15] Until BellSouth can just justify the legitimacy of a standard error that is twice the size of those accepted in similar workshops throughout the country, its test procedure and results are simply not credible and should not be adopted by this Commission.

Finally, BellSouth views the appropriate testing procedure to be the conduct of a two-tailed test at the 0.05 level of significance. In his summary, Dr. Scheurer, of the BellSouth team, considered this issue as "ready to call." MCI begs to differ. Tests of parity are performed to determine whether the provisions of the 1996 Telecommunications Act have been met. The law is clear; the ILECs must provide service to the CLECs of a quality <u>at least equal to</u> that which it provides itself. Thus, it is perfectly legal for BellSouth to provide the CLECs with higher quality service than it provides itself; the law is violated when BellSouth provides the CLECs with lower quality service. This makes any statistical test of whether this law has been violated a one-tailed test -- without debate. (MCI does not doubt that there is managerial information of importance to BellSouth if a two-tailed test indicated that it was

[15] <u>See</u> Point 5 in Dr. Mallows' LPSC statistics workshop comments.

supplying higher quality services to the CLECs than to itself, but that type of managerial information should be of no concern to those concerned with enforcing the law.)

The importance of this point cannot be overemphasized. Consider the BellSouth discussion of Maintenance Average Duration on page 25 of its Interim Statistical Analysis. In discussing the drop of test statistic values from a -3 to -6 range to a -1.91 to -1.93 range as a result of adjusting the BellSouth data, BellSouth notes that the indication of non-parity "is not at all strong after the data are adjusted." There are roughly three chances in 100 of getting a test statistic value as extreme as any one of those found for the adjusted data. MCI contends that most statisticians would consider such result as fairly strong evidence of non-parity. How, then, could BellSouth justify its conclusion? BellSouth undertakes two-tailed tests at 5% significance level. This makes the critical regions on either tail 2.5%. Since 3% is greater than 2.5%, BellSouth does not reject parity.

If the test is changed to a one-tailed test, as it should be, all of the rejection probability is put in the lower tail of the distribution. Since 3% is less than the 5% probability of rejection, parity is rejected. Splitting the probability of rejection, as is done in a two-tailed test, is inappropriate, and could result in incorrect inferences regarding the service BellSouth is providing to CLECs.

D. Problems With The BellSouth Criticism of the LCUG/FCC Methodology

The BellSouth analysts' rejection of the FCC/LCUG testing procedures rests solely on their contention that these procedures require some strenuous assumptions that are not satisfied by their rote application. MCI is mystified by this argument. [The LCUG/FCC tests require that the conditions of the Central Limit Theorem be met in order for those statistics to be normally distributed.] But, as noted above, so does the BellSouth procedure.[16] Since both procedures are based on the same data, it hardly seems likely that the LCUG/FCC approach fails to satisfy some assumptions that the BellSouth approach satisfies. The difference in the procedures begins when BellSouth randomly assigns wire center mean differences to various groups. While this procedure may remove some geographic dependency in the data, when BellSouth computes group means and the between group variance, its computations are based on precisely the same data as would be used in the LCUG/FCC analyses. Thus, the BellSouth claim, made in Appendix B, that its procedure satisfies assumptions that the LCUG/FCC procedures do not, is singularly opaque. Furthermore, as Dr. Mallows notes in point 4 of his statistics workshop comments, "[t]he choice of statistic does not depend on any assumptions; though of course the efficacy of the resulting procedures will depend on how the data actually behave." This suggests that the whole BellSouth argument about assumptions is a smokescreen to allow it to claim legitimacy for a procedure that generally results in cutting the appropriate test statistic values in half.

The assumption that BellSouth seems most concerned that the LCUG/FCC approach violates is one of "independence." MCI is unclear what BellSouth means by this term,

[16] See Interim Statistical Analysis, p. B 10.

because "independence" has several alternative connotations in statistical parlance (e.g., independent variables, stochastic independence, serial independence, etc.). At any rate, BellSouth spends considerable time on pages 23 and 24 of its Interim Statistical Analysis attempting to determine whether the "assumption of independence between the observations" (p.23) is satisfied. BellSouth does this by fitting lines to scatter plots of BellSouth wire center means for Maintenance Average Duration against corresponding CLEC wire center means for both adjusted and unadjusted data (Figures 20 and 21). If the observations were independent, BellSouth argues, the lines should be flat (p. 24). BellSouth concludes that upward sloping indicates the observations are not independent. Consequently, BellSouth contends the crucial assumption of the LCUG/FCC approach is not met.

This analysis is totally without merit. First, in a world of parity, a line fitted to a scatter of BellSouth wire center means against corresponding CLEC wire center means should have a slope of unity, not zero. If CLEC Maintenance Average Duration increases by an hour, for any reason, so should BellSouth's. If it increases by less, BellSouth is discriminating. Since the BellSouth analysts found slopes of 0.25 and 0.30 for the unadjusted and adjusted data, respectively, they have in fact provided weak evidence of BellSouth discrimination in the provision of Maintenance Average Duration (weak, because the regression is likely specified incorrectly).

Second, regardless of what the regressions do demonstrate, they definitely do not demonstrate a violation of the "assumption of independence between the observations."

Consider this proposition: When one estimates a regression equation, one assumes that observations on the dependent variable are independent realizations of an n-variate random variable (see, for example, Theil, 1970). When one eventually fits the model and finds that some explanatory variable explains X% of the variation in the dependent variable, does this mean that the "assumption of independence between the observations" is violated? Of course not. If it did, one could never legitimately estimate a regression because the regression would always be in violation of one of the assumptions. *The point is that correlation between variables has nothing to do with independence between the observations on a particular variable.*

Interpreting the BellSouth analysis of assumption violation in its most favorable light, it appears that BellSouth is suggesting that the reason that the slope of the regression line diverges from unity is that the LCUG approach has not taken into account a number of confounding factors that systematically affect the relationship between the BellSouth and CLEC wire center means. If so, BellSouth would have an excellent point. But, the BellSouth approach is neither the only way, nor the best way, to handle this problem. Dr. Mallows, in the latter stages of point 4 of his statistics workshop comments, indicates how the LCUG/FCC approach could be used to handle the problem of confounding variables.

MCI proposes two additional alternatives, both of which have the ability to factor out the effects of all relevant confounding variables, and both of which provide only one relevant test statistic. The first approach is to estimate (the appropriate form of) the model posited in Section A.3 above. To reiterate, this model amounts to pooling the SQM data for BellSouth and the CLECs and estimating a regression equation consisting of a constant term, a group dummy (say, d=0 for BellSouth observations, d=1 for CLEC observations), analogously defined dummies for the various sub-categories of Order Completion Interval characteristics, a set of 229 wire center dummies (or a set of dummies for a more appropriate geographical categorization), and a set of interaction variables created by multiplying the group dummy times all other included variables. Within this context, the t statistic on the group dummy coefficient amounts to a *ceteris paribus* means difference test with all of the confounding influences factored out; an F test on the joint significance of the group dummy and all interaction coefficients provides a *mutatis mutandis* test. The second alternative is more in the spirit of the FCC/LCUG approach; in particular, it incorporates the LCUG sensitivity to departures from equality of variance. In this approach, MCI estimates the above model excluding the constant term, group dummy and interactions. The residuals of this model can be viewed as the SQM with all of the effects of the confounding variables filtered out. One could then legitimately perform the LCUG z-test on this purged data. This appears to be the best approach to handling the problems illuminated by the BellSouth study without encountering even more pernicious problems inherent in the BellSouth approach.

III. CONCLUSION

MCI encourages the LPSC to adopt the modified z-test and permutation test as the appropriate test statistics to determine parity. These test statistics provide optimal power for detecting the types of departures from parity that prevent CLECs from competing on equal terms. Further, concerns regarding sample sizes and confounding factors are resolvable using the modified z-test and permutation test. It is no coincidence that US West, BellAtlantic, GTE, Nevada Bell, and Pacific Bell have all agreed to the use of the modified z-test to determine parity.

On the other hand, BellSouth's proposed test is not sensitive to differences in variation, can be used to cover-up significant differences in means, and uses subjective analysis to determine whether parity exists. Furthermore, the BellSouth proposal does not yield swift results, a must in the newly emerging competitive marketplace. In sum, BellSouth's proposed test is simply not appropriate for determining parity and should be rejected by this Commission.

Submitted by:

Katherine W. King
Gordon D. Polozola
KEAN, MILLER, HAWTHORNE,
D'ARMOND, McCOWAN & JARMAN, L.L.P.
P. O. Box 3513
Baton Rouge, LA 70821

Martha McMillin
MCI TELECOMMUNICATIONS
CORPORATION
780 Johnson Ferry Road, Suite 700
Atlanta, Georgia 30342

Attorneys for MCI Telecommunications
Corporation

CERTIFICATE OF SERVICE

I hereby certify that a copy of the above and foregoing has this date been served via hand delivery, electronic mail, facsimile and/or overnight mail to all persons on the Official Service List.

Baton Rouge, Louisiana, this 15th day of December, 1998.

Gordon D. Polozola

Using Permutation Tests to Evaluate the Significance of CLEC vs. ILEC Service Quality Differentials

JOHN D. JACKSON

E-GROUP, phone 334.844.2926, 334.844.2615.fax, jjackson@business.auburn.edu.

ABSTRACT (and illustration) OF THE CENTRAL ISSUE

The purpose of the LCUG-Z test is to evaluate whether or not an ILEC is providing services to itself and to CLECs on an equal basis. Using the LCUG-Z to test for parity is only valid for large samples since the test is based on the assumption that the underlying samples for both the CLEC and ILEC are normally distributed (the histogram looks like the bell curve). While the number of ILEC observations typically is large enough to invoke the Central Limit Theorem, CLEC sample sizes are often too small for the theorem to apply and, as a consequence, the distribution of the CLEC sample means is unknown. In cases where the sample size for the CLEC is very small, the permutations test is required since this test does not require one to know *a priori* the distribution of the CLEC sample.

The Permutation test is conducted as follows. Assume you have a CLEC sample of 10 observations and an ILEC sample of 90 observations so that the total sample size is 100 observations. Let's say this sample is of order-times. The first step is to calculate the simple LCUG-Z for these two samples. We cannot, however, make conclusions from this Z value in the same manner we could for large samples. For example, if the LCUG-Z is calculated to be 2.5, we would certainly conclude that there is a lack of parity if the CLEC sample is large, but we should be hesitant to make that conclusion for small samples.

In order to make use of the LCUG-Z for small samples, we need two additional steps. The second step of the Permutations test is to pool the two samples (100 observations). It might help to think of putting 100 index cards (numbered 1 through 100) with an order-time written on each one into a bowl. From this bowl, we will draw 10 cards and calculate the mean and standard deviation for those 10 order-times. Then, we will take the remaining 90 cards and calculate the mean and standard deviation for those order-times. Using these means and deviations, we calculate a Z value using the LCUG-Z formula. After getting this Z value, all the cards are put back into the bowl and we draw 10 cards again, calculating the means and standard deviations for both the drawn cards and the remaining 90 cards. Again, the Z test is calculated. If we happen to get the same 10 cards as from an earlier draw (recall, the cards are numbered), we put the cards back and draw again. All of our 10 card samples must be a unique set of 10 cards although a specific order-time, say 10 seconds, might appear on a number of cards. Given the full sample size of 100 observations and the CLEC sample of 10 observations, we can repeat this process 17.3 trillion times [100!/(10!)(90!)], getting 17.3 trillion Z values. Rather than

spend the rest of our lives doing these calculations, we instead draw a sample from the 17.3 trillion potential samples of (perhaps) 1,000 samples. These 1,000 samples provide us with 1,000 Z values.

Armed with our initial, simple LCUG-Z of 2.5 and a sample of 1,000 Z values from our drawing procedure, we are prepared to reach a conclusion about parity. In step three of our algorithm, we order the 1,000 Z values from lowest to highest and then find where our LCUG-Z value of 2.5 falls in this ordering. If only 10% of the 1,000 Z values are greater than 2.5, then we can reject parity at the 10% significance level. Likewise, if only 5% of the 1,000 Z values lie above 2.5, then we can reject parity at the 5% level.

Using Permutation Tests to Evaluate the Significance of CLEC vs. ILEC Service Quality Differentials

JOHN D. JACKSON
Department of Economics, 203 Lowder Business Bldg., Auburn University, AL 36849, 334.844.2926, 334.844.4615.fax, jjackson@business.auburn.edu.

I. Introduction

The problem that confronts us, statisticians and laymen alike, is to determine whether the ILEC provides the same quality of a particular service to a given CLEC that it provides to itself. In order to understand the role and the mechanics of permutation testing in this decision, a number of prefatory points must be made, and a step by step detailing of the application is required. The explanation will proceed as follows: First, the need for permutation testing must be justified. This requires an understanding of the theoretical foundations of the LCUG-Z statistic, why it is suspect in small samples, and why the permutation testing procedure is the appropriate means of solving the small sample problem. Second, given that permutation tests are to be used, how are they to be conducted? An explanation of permutation tests in a means difference framework, closely paralleling Dr. Colin Mallows' circulating document, is presented. An illustration of the procedure using a sample small enough for the full permutation distribution to be available is provided, and the modifications inherent in any practical application of this principle are detailed. Finally, since users of this proposed methodology will presumably need to translate it into a computer program that analyzes real data, I conclude with a presentation of the logic underlying any computer program attempting to conduct this type of hypothesis test. In addition, I provide two programs (one previously circulated by Dr. Mallows and one a SAS program) that will perform the required operations (efficiently) for intermediate samples and (eventually) for large samples.

II. What's Wrong With the LCUG-Z when the CLECs: Don't Have "Enough" Observations on a Particular Service Quality Measure

Let us begin with a review of the statistical issues involved. Consider a particular service typically provided by an ILEC to a CLEC, say, providing service change orders. Our interest is in whether BANY provides this service to the CLEC as quickly, from order receipt to completion of work, as it does for itself. Clearly, we cannot simply look at only one illustrative example of how fast BANY processes an order for itself and compare it to a corresponding illustrative example of their processing of a CLEC's order. Not all orders are alike; each is affected by innumerable and unknowable factors. We may be able to identify and control for the effects of some of these factors, but it is simply not possible to do so for all of the factors. Thus, viewing only one particular order as "representative" could be highly misleading.

Put another way, we can treat the length of time it takes to process a service change order as if it were a random variable whose values follow some statistical distribution. That is, there is some probability associated with the observation of any specific order-time value. The weighted average of all possible order-time outcomes, where the weights are the probability that the corresponding outcome occurs, is called the "expected value" or "mean" of the distribution, and is usually denoted by the Greek letter μ. It measures the location of the distribution in the real number line; i.e., it tells us where the center of the distribution is located and is hence sometimes referred to as a measure of central tendency. Other important parameters that help us characterize the distribution of order-intervals are the variance (usually denoted by the Greek letter σ^2) of the distribution, and its square root, the standard deviation (denoted by σ). These parameters measure the dispersion of the outcomes about their mean value: the larger σ^2 or σ, the more spread out the distribution, the smaller σ^2 or σ, the more compact the distribution. It should be emphasized that, as a rule, we do not know the true mean, the true variance, or the shape of the distribution.

Applied to the parity question, we are faced with some distribution of times taken by BANY to execute a service change order for their own customers, characterized by a true (but unknown) mean (μ_{ILEC}) and a true (but unknown) variance (σ^2_{ILEC}). Likewise, we have a distribution of times taken by BANY to execute a service change order for a competitive local exchange carrier's customers, characterized by a possibly different, but equally unknown, mean (μ_{LEC}) and variance (σ^2_{CLEC}). From a statistical perspective, then, the question of service quality parity for this particular measure boils down to the question of whether the mean of the BANY distribution of service change order times differs significantly from CLEC service change order times.[1] Answering this question requires the test of the "null hypothesis" H_0: $\mu_{ILEC} = \mu_{CLEC}$ (which is assumed to be true when computing the test statistic) against the "alternative hypothesis" H_1: $\mu_{ILEC} < \mu_{CLEC.}$

Conducting this hypothesis test requires first that we have random samples of observations from both the CLEC and ILEC distributions of service change order times. Suppose we draw n_{ILEC} observations (X_i , $i = 1, \ldots, n_{ILEC}$) from the BANY distribution and compute the sample mean

$$\overline{X}_{ILEC} \quad (= \sum_{i=1}^{n_{ILEC}} X_i / n_{ILEC})$$

and variance

$$S^2 \quad [= \sum_{i=1}^{n_{ILEC}} (X_i - \overline{X}_{ILEC})^2 / (n_{ILEC} - 1)]$$

of those order-times. Similarly we draw n $_{CLEC}$ observations (X_i , $i = 1, \ldots, n_{CLEC}$) from the CLEC distribution and compute the sample mean

$$\overline{X}_{CLEC} \quad (= \sum_{i=1}^{n_{CLEC}} X_i / n_{CLEC})$$

and variance

$$S^2 \quad [= \sum_{i=1}^{n_{CLEC}} (X_i - \overline{X}_{CLEC})^2 / (n_{CLEC} - 1)]$$

of their order-times. These sample statistics are ("best" point) estimates of their population counterparts.

[1] This is really boiling the decision down quite a bit. A true test of parity is a test of whether the ILEC and CLEC distributions are the same. Thus parity requires not only that the means be the same, but also the variances and higher moments be the same as well. The means difference tests that we conduct implicitly assumes that these other requirements are met.

Here is where the LCUG-Z comes in. In order to test the above null hypothesis (i.e., H_0: $\mu_{ILEC} = \mu_{CLEC}$) we must develop a test statistic that meets certain requirements: It must be based on a random variable that is a function of the parameters being tested (i.e., μ_{ILEC} and μ_{CLEC}). It must be based on estimators (i.e., $\overline{X}_{ILEC}$ and $\overline{X}_{CLEC}$) of the parameters being tested. *It must follow a known distribution* (in this case, a standard normal). We must be able to compute a numerical value for it under the assumption that H_0 is true. A number of statistics meet this requirement, but it has been decided that the LCUG-Z is the one that will be used in the service quality parity analysis. Thus we will confine our attention to it in the following.

In its purest form, the LCUG-Z can be written based on the following normally distributed random variable

$$Z = \frac{(\overline{X}_{CLEC} - \overline{X}_{ILEC}) - (\mu_{CLEC} - \mu_{ILEC})}{\sqrt{\left(\frac{\sigma^2_{CLEC}}{n_{CLEC}} + \frac{\sigma^2_{ILEC}}{n_{ILEC}}\right)}}. \tag{1}$$

Under the assumptions that the null hypothesis is true (i.e., $\mu_{ILEC} = \mu_{CLEC}$), that $\sigma^2_{ILEC} = \sigma^2_{CLEC}$, and that this common variance is estimated by S^2_{ILEC}, it attains its more familiar form as a statistic[2]

$$Z = \frac{\overline{X}_{CLEC} - \overline{X}_{ILEC}}{S_{ILEC}\sqrt{\left(\frac{1}{n_{CLEC}} + \frac{1}{n_{ILEC}}\right)}}. \tag{2}$$

This statistic is known (we will come back to how we know this momentarily) to follow a standard normal distribution (basically, this means that the distribution looks like a bell curve with mean = 0 and variance = 1). Equation 2 further indicates that this Z statistic is readily computable since it depends only on statistics ($\overline{X}$, S, and n) computed from the order-time samples discussed above.

From this point on, conducting the test is simple: Compute the value for Z using equation (2), call it Z*. Select a level of significance appropriate to the test

[2] We substitute the ILEC variance for the CLEC variance in order to improve the sensitivity of the statistic to situations in which the estimated CLEC variance exceeds the estimated ILEC variance.

(α = 0.10,α = 0.05, and α = 0.01 are traditional, but not sacrosanct) which in turn determines a critical value of Z (Z_C). The decision of whether or not to reject the null hypothesis is based on a comparison of the computed and critical values of Z If $Z^* > Z_c$, reject H_0 implying parity in order-time service provision is not present; if $Z^* < Z_C$, the hypothesis of parity in service provision cannot be rejected at the chosen level of significance.

The preceding discussion provides a general framework for analyzing potential problems arising from the use of the LCUG-Z parity hypotheses. In the remainder of this section, we address a number of these problems as they relate to questions deriving from small CLEC sample sizes. Specifically, we consider why the LCUG-Z statistic is suspect for small sample sizes, why aggregating across CLECs or over time is an unsatisfactory solution to the small sample size problem, and why the techniques of permutation testing provide a satisfactory solution to the problem.

WHY IS THE LCUG-Z SUSPECT FOR SMALL SAMPLE SIZES?

To answer this question, we must consider the theoretical foundations of the LCUG-Z -- where it came from and why it follows a standard normal distribution. There are at least three theoretical cornerstones to the LCUG-Z: the *Central Limit Theorem,* a result on the difference between two normal distributions, and a result on standardizing normal distributions. We will briefly outline each of these and then combine them to throw some light on the small sample problem.

The Central Limit Theorem (CLT) presents what is perhaps the most powerful result of inferential statistics. Basically, it states that the distribution of sample means is approximately normal with mean equal to the population mean and variance equal to the population variance divided by the sample size. While these results hold for "large" sample sizes (and nobody knows how large is "large"), it also holds *regardless of the distribution of the parent population from which the sample is drawn.* In our example this means that we do not need to know what the distribution of times required to execute a service change order looks like for BANY, nor do we need to know what the corresponding CLEC's distribution looks like. All we need to know is that we have computed means from random samples of each of these two distributions. The theorem guarantees us that each mean is drawn from an approximately normal distribution whose (true, but unknown) mean is the mean of the corresponding population and whose (true, but unknown) variance is the variance of the corresponding population divided by the relevant sample size, assuming that each of our samples is large enough.[3] Put another way, the theorem tells us that $\overline{X}_{ILEC}$ follows an approximately normal distribution with mean μ_{ILEC} and variance σ^2_{ILEC}/n_{ILEC}, and similarly that $\overline{X}_{CLEC}$ follows an approximately normal distribution with mean μ_{CLEC} and variance σ^2_{CLEC}/n_{CLEC} – the approximations will be close assuming that n_{ILEC} and n_{CLEC}, respectively, are sufficiently large.

A second step in understanding the basis of the LCUG-Z lies in a result from statistical distribution theory. Specifically, it can be shown that if we create a new statistic by taking the difference between two independent normally distributed random variables, that new statistic will also be normally distributed with mean equal to the difference between the means of the two normal random variables and variance equal to the sum of their variances. Thus, since $\overline{X}_{ILEC}$ follows a normal distribution with mean μ_{ILEC} and variance σ^2_{ILEC}/n_{ILEC}, and since $\overline{X}_{CLEC}$ follows a normal distribution with mean μ_{CLEC} and variance σ^2_{CLEC}/n_{CLEC}, $(\overline{X}_{ILEC} - \overline{X}_{CLEC})$ follows a normal distribution with mean $(\mu_{ILEC} - \mu_{CLEC})$ and variance $[(\sigma^2_{ILEC}/n_{ILEC}) + (\sigma^2_{CLEC}/n_{CLEC})]$.

Finally, it is known that any normally distributed random variable can be converted into one following a standard normal (a normal distribution whose mean is zero and whose variance is one), by subtracting out its mean and dividing through by its standard deviation. Performing this standardization operation on the above means-differenced random variable leads to

$$Z = \frac{(\overline{X}_{CLEC} - \overline{X}_{ILEC}) - (\mu_{CLEC} - \mu_{ILEC})}{\sqrt{\left(\frac{\sigma^2_{CLEC}}{n_{CLEC}} + \frac{\sigma^2_{ILEC}}{n_{ILEC}}\right)}}. \qquad (3)$$

which is the random variable expressed in equation (1) above upon which the LCUG Z is based.

The above discussion has been overly tutorial, but hopefully with a purpose. The intent has not been to numb the reader's brain with statistical trivia. Rather

[3] Be clear. the distribution that the CLT refers to is the distribution of sample means -- not order times. That is, we could draw one sample of ILEC order times and compute its mean, we could draw another sample and compute its -- undoubtedly different-- mean, we could draw a third sample .... These means that we could compute follow a statistical distribution, and it is <u>this</u> distribution that the CLT shows to be asymptotically normal.

we have attempted to highlight precisely why the LCUG-Z is suspect in the case of small CLEC samples. The problem is not how reliably $\overline{X}_{CLEC}$ estimates μ_{CLEC} or S^2_{CLEC} estimates σ^2_{CLEC}; if these statistics are computed using the formulae given above, they are "best" estimates of their corresponding parameters -- for any given sample size. The problem is that if CLEC samples are small, there is no guarantee that $\overline{X}_{CLEC}$ follows a normal distribution with mean μ_{CLEC} and variance $\sigma^2_{CLEC} / n_{CLEC}$, so that there is no guarantee that $(\overline{X}_{ILEC} - \overline{X}_{CLEC})$ follows a normal distribution with mean $(\mu_{ILEC} - \mu_{CLEC})$ and variance $[(\sigma^2_{ILEC}/n_{ILEC}) + (\sigma^2_{CLEC}/n_{CLEC})]$, and consequently, the LCUG-Z may not follow a standard normal distribution. Instead, it may follow some unknown distribution.[4] Clearly, if we base our decisions on parity in service quality on the standard normal distribution when the LCUG-Z in fact follows some other unknown distribution, we run the risk of making erroneous inferences that could cost all agents involved considerably.

III. Permutation Testing as a Solution to the Small Sample Problem

In a recently circulating document, Dr. Colin Mallows has referred to a number of alternative nonparametric procedures, including sign-rank tests and bootstrap methods, to address our difficulty. Perhaps the most promising potential solution is permutation testing. What is permutation testing, anyway? And how does it solve the small sample problem?

Recall that the problem caused by small CLEC sample sizes is that we do not know what distribution the LCUG-Z follows. In its simplest form, a permutation testing procedure solves this problem by literally computing an empirical distribution for the statistic based all possible samples of size n_{CLEC} that could possibly be drawn from a pooled sample of $n_{ILEC} + n_{CLEC}$. We can then compare the value of Z computed from equation 2 using the actual CLEC and ILEC samples drawn to this empirical distribution to determine the probability of finding a Z value larger than this computed value. We can then use this probability directly to decide whether to accept or reject the parity hypothesis, or

[4] The above discussion glosses over some substantive issues relating to the shape of the distribution of the underlying population. (1) normality: if the population is normally distributed, small samples are not a problem; the statistic will be normally distributed regardless of sample size. Alternatively, if the population is not normally distributed, the distribution of sample means may not be very close to normal, even for samples as large as 30. The effect of this is that the distribution of Z can be far from standard normal. (2) Symmetry: If the populations are skewed, as are several SQMs that we must consider, and we are dealing with one-sided tests, the normal approximation can be quite bad even for samples as large as thirty.

we can convert this probability into a "corrected" Z statistic which, unlike the original Z, will produce exactly the correct type 1 error for the test. While the intuition underlying this approach is straightforward, the approach itself is not often used and it deserves considerably more explanation. The point to be made here is that a permutations testing approach to testing the parity hypothesis (as measured by differences in mean service qualities) can produce an exact (rather than approximate) test even if CLEC samples are quite small. The following section considers in some detail the "how and why" of this result.

1. PERMUTATION TESTS OF MEANS DIFFERENCE HYPOTHESES

The above discussion points out the fact that permutation methods can produce an exact test of the parity hypothesis even when CLEC samples are small.[5] This explains why we would be interested in such procedures, but it does little to explain how these procedures work. This section addresses this question. First, we consider the conceptual framework of permutations methods applied to testing for service quality parity. Then, we illustrate this procedure using a sample small enough for the full permutation distribution to be available. Next, we consider the modifications required when dealing with real world data and sample sizes. Finally we illustrate the application of these procedures, first using a sample of hypothetical data ($n_{CLEC} = 11$, $n_{ILEC} = 214$) provided by BANY, and then using a sample of California data on an actual service quality measure ($n_{ILEC} = 167533$ and n_{CLEC} varies from 1-337, depending in the firm).

The general idea behind permutation tests is to use information on every possible sample of a given size that could be drawn from a particular "population" (where the number of observations in the CLEC and ILEC samples determine both the sample size and the population size) to derive an empirical distribution, called the permutation distribution, of a specific test statistic. We can then compare the value of the test statistic computed for the sample actually drawn to this distribution to make inferences leading to the acceptance or rejection of a statistical hypothesis. Clearly, the first question that arises is how many samples should we deal with and what do we do with them to generate this permutation distribution?

[5] We emphasize small CLEC samples because that is the particular problem that we are concerned with at the moment. Permutation testing is perfectly general; it does not require either group to have a small sample size. If, however, neither group has a small sample size, there are a number of equally valid and less tedious procedures available for testing the same hypothesis (e.g., the LCUG-Z).

Recall that the hypothesis that we wish to test is H_0: $\mu_{ILEC} = \mu_{CLEC}$. As before, we construct the test assuming that parity holds. In particular, we note that under the parity hypothesis, there should be no systematic differences between the CLEC and ILEC samples. This is the only assumption of the permutation approach. Thus if we pool both samples to form a "population" of size $n_{TOT} = n_{ILEC} + n_{CLEC}$ and then randomly allocate any n_{CLEC} of the n_{TOT} observations to one group and the remaining $n_{TO\ T} - n_{CLEC} = n_{ILEC}$ observations to a second group, each such allocation should be equally likely. They can thus be viewed as random samples of size n_{CLEC} and n_{ILEC}, respectively, from a population of size n_{TOT}. The total number of possible independent random samples of size n_{CLEC} that can be drawn from a population of size n_{TOT}, where the ordering of the observations makes no difference, is given by the counting rule $N = [n_{TOT}!/(n_{CLEC}!)(n_{ILEC}!)]$.[6] Since drawing any sample of size n_{CLEC} uniquely determines a corresponding sample of size n_{ILEC} , this is also the number of samples of size n_{ILEC} that can be drawn. To be clear, we have N pair of random samples, each pair containing one sample of size n_{CLEC} and one of size n_{ILEC} .

Once we select a pair of samples, we compute the mean of each sample and the standard deviation of the larger ILEC sample. Using these data, we apply equation (2) to compute the LCUG-Z for that sample pair. We repeat this process for each of the N pairs of samples. The resulting set of Z statistics includes every value that the LCUG-Z could possibly take on, given the initial "population" of observations. The (frequency) distribution of these values is the permutation distribution.

Next, we compare the LCUG-Z value computed for the actual samples we observed to this empirical distribution of Z values. Our intent is to determine the percent of Z values in the distribution lying above the computed value. This can be viewed as the probability if committing a type 1 error in testing our parity hypothesis, since if we reject, this tells us the probability of rejecting a true null. (Clearly it is possible to compute Z values larger than the one observed; the key characteristic of a permutation test is that this percentage is a precise measure of how many - at least as a percent of $N-Z$ values would be computed above that the observed value, given the population under study.) Another way to view that percentage is as the true p-value of the test. That is, it tells us the probability

[6] In general, there are (n_{TOT}) raised to the n_{CLEC} power different ways of selecting samples on size n_{CLEC}. If we sample without replacement, this number drops to $n_{TOT}!/(n_{TOT} - n_{CLEC})!$ Finally, if order makes no difference (i.e., if sample 1 selects observations 1,2,and 3 and sample 2 selects observations 3,1,and 2, then sample 1 and sample 2 are viewed as the same sample) the number falls still further to $n_{TOT}!/[(n_{TOT} - n_{CLEC})!(n_{CLEC})!]$ as noted above.

of observing a Z value greater than or equal to the one in question, given the population under study.

From equation 2, it should be obvious that the larger the difference in the two means the larger the Z statistic, and hence the smaller the percent of the permutation distribution lying above it, *ceteris paribus*. This means that if we wish, we can stop right here. If there is only a small proportion of values lying above the computed Z, the probability of getting a Z this large by chance is small (the type I error probability is low), indicating a larger difference between the two mean service quality measures, and further implying a rejection of the service parity hypothesis. If we wish to go further, we can "back-out" a "corrected Z score" from this p-value by simply asking what Z value would cut off that much probability in the right side tail of a standard normal distribution. This "corrected" Z value is preferable to the LCUG-Z for the observed samples because the former provides a test statistic with an exactly determined type 1 error. A simple example is offered to illustrate these procedures.

2. AN ILLUSTRATION

Suppose we have a very small sample of data of order-time data for a CLEC. Let this sample be (3,5) so that $n_{CLEC} = 2$. Also suppose that we have a very small sample of order-time data for the ILEC. Let this sample be (1,2,4) so that $n_{ILEC} = 3$. Now we pool these two samples to give us a "population" of size $n_{TOT} = 5$, consisting of the observations (1, 2, 3, 4, 5). Using our counting rule, there are $[5!/(2!)(3!)] = 10$ different samples of size 2 (having 10 corresponding samples of size 3) that can be drawn from this population of size 5. For a problem this small, it is easy to write out exactly what these 10 sample pairs are. Remember, duplicate observations such as (1, 1), (2, 2), ... (5, 5) are not possible because we are sampling without replacement. Also recall that samples like (2, 1) are treated as identical to samples such as (1, 2) because the ordering of the observations within the sample makes no difference. The 10 sample pairs and the computations relating to each are presented in Table I below.

Table I: Finding the Permutation Distribution of Z Values

Sample Number	CLEC Sample	CLEC Mean	ILEC Sample	ILEC Mean	ILEC S.D.	S.D. of Mean	Means Diff	Z Stat
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	1, 2	1.5	3, 4, 5	4.0	1	0.91287	-2.50	-2.74
2	1, 3	2.0	2, 4, 5	3.67	1.528	1.3944	-1.67	-1.20
3	1, 4	2.5	2, 3, 5	3.33	1.528	1.3944	-0.83	-0.60
4	1, 5	3.0	2, 3, 4	3.0	1	0.91287	0.0	0.0
5	2, 3	2.5	1, 4, 5	3.33	2.082	1.900	-0.83	-0.44
6	2, 4	3.0	1, 3, 5	3.0	2	1.826	0.0	0.0
7	2, 5	3.5	1, 3, 4	2.67	1.528	1.3944	0.83	0.60
8	3, 4	3.5	1, 2, 5	2.67	2.082	1.900	0.83	0.44
9	3, 5	4.0	1, 2, 4	2.33	1.528	1.3944	1.67	1.20
10	4, 5	4.5	1, 2, 3	2.0	1	0.91287	2.50	2.74

The CLEC sample is given in column 2 and the corresponding ILEC sample in column 4. In order to compute the LCUG-Z for each sample pair according to equation 2, we must have information as to the CLEC sample mean (col.3), the ILEC sample mean (col. 5), the ILEC standard deviation (col.6), and sample size information $\{[(1/n_{CLEC}) + (1/n_{ILEC})]^{.5} = (5/6)^{.5} = 0.91287\}$. Column 7 is the denominator of equation (2); column 8 is the numerator of equation (2); and column 9 is their ratio, the value of the LCUG-Z computed for that sample pair. If we arrange the values of Z in ascending order, we have the permutation distribution for this problem, vis.,

-2.74, -1.20, -.60, -.44, 0, 0, .44, .60, 1.20, 2.74.

Now let us compute the LCUG-Z for the observed sample

$$Z = \frac{\overline{X}_{CLEC} - \overline{X}_{ILEC}}{S_{ILEC}\sqrt{\left(\frac{1}{n_{CLEC}} + \frac{1}{n_{ILEC}}\right)}} = \frac{4.0 - 2.33}{1.528\sqrt{\left(\frac{1}{2} + \frac{1}{3}\right)}} = \frac{1.67}{1.528 \cdot 0.91287} = 1.2 .$$

We observe that one of the values in the permutation distribution lie above this computed value, and one is equal to it. For a small number of permutation samples, such as we have here, this equality is a matter of some concern. This problem of "ties" is typically handled by treating the LCUG Z as if it ranked 8.5 (=9-0.5) rather than 9th in the permutation sample hierarchy.[7] Thus 85% [=(9-

[7] This type of adjustment will make no appreciable difference when the number of permutation samples is large. Additionally, ties are simply unlikely in this case.

0.5)/10] of the possible Z values less than the computed LCUG Z and 15% greater. This means that the null hypothesis of parity can be marginally rejected at the 15% level of significance. Alternatively, we note that a Z value of 1.03 (found in a standard normal table) cuts off 15% of the probability in the right tail of a standard normal distribution. This means that if we use the "corrected" Z value to conduct the standard test, we will still marginally reject the null of parity in service provision at the 15% level of significance.

3. SOME MODIFICATIONS OF THE PROCEDURES REQUIRED TO HANDLE MORE REALISTIC PROBLEMS

The above illustration was quite small. It is not often that practical problems allow us to write out the full permutation distribution. For example, consider the hypothetical data set furnished by BANY to get some notion of the practicability of the permutation test. We have 214 hypothetical observations on the ILEC and 11 observations on the CLEC for a total of 225 observations in the pooled sample. This is still a very small sample by ILEC standards for some service quality measures. Nevertheless, our counting rule tells us that there are roughly $1.5\text{x}10^{18}$ different samples of size 11 that could be drawn a population of size 225. One can imagine several medieval monks spending their entire lives trying to write out the full permutation distribution for this problem, and still not finishing the job.

The message is clear: for even moderate sample sizes, it is simply not possible to analyze the full permutation distribution. Consequently we must sample from that distribution and base our inferences on that sample. At least in this case we control the number of sample pairs to be analyzed. For example, we could randomly select 1000 sample pairs and compute the p-value and "corrected" Z value just as we outlined above. Of course we would need a computer to conduct this analysis, because conducting 1000 means difference tests with samples of 11 and 214 could be quite time consuming if done by hand. We could then select another random seed, draw another 1000 sample pairs, and compute the p-value and "corrected" Z for this data set. Repeating this experiment several times should produce not only a good estimate of the p-value and corrected Z when the results are combined, but the separate analyses should also provide an indication of the robustness of these estimates to perturbations in the permutation sample selected.

Now let us turn to some illustrations of the modified procedure applied to more realistic problems.

4. EXAMPLES OF THE PROCEDURE APPLIED IN MORE REALISTIC SETTINGS:

Dr. Colin Mallows has written a computer program using the S system to conduct the type of tests discussed above. He has described (various forms of) it in some detail in a series of e-mail messages and reports. Also, we here at **E-Group** have written a general SAS program to perform the same types of analyses. Below we present the results of each program applied to the BANY hypothetical data discussed above.

First note that if we compute the LCUG-Z for the data provided by BANY, we find the Z value is 1.39 and the *p*-value is about 0.083. These results obtain regardless of which program is used. However once we start pulling random samples from the permutation distribution, we will observe program differences if for no reason other than the fact that the two programs are using different samples.

Dr. Mallows selected four sets of 1000 randomly selected sample pairs (n_{CLEC} = 11, n_{ILEC} = 214) and found the following results: 0.0905(1.284), 0.1085(1.235), 0.1005(1.279), 0.1015(1.273), and combined 0.10(1.282)where the first number given is the p-value and the second , in parenthesis, is the corrected Z. The final two numbers are the results implied by combining the results for the four separate runs. We conducted the same experiment using the same BANY data but using our SAS program and found the following: 0.1075(1.24), 0.0915(1.332), 0.0855(1.369), 0.0985(1.29), and combined 0.09525 (1.308). Note that both programs produce very similar results and that both indicate that, in this particular illustration, the rote application of the LCUG-Z indicates a greater likelihood of rejecting parity than it should (every corrected Z is smaller than the computed Z). It is worth noting that both programs were able to complete these calculations in a matter of only a few seconds.

Since BANY's set of hypothetical data is still somewhat modest in size, we used our SAS program to apply these procedures to a set of data on "receipt to clear in minutes" for California CLECs and ILECs provided by Jim Kistner. In this data set, n_{ILEC} = 167533 and n_{CLEC} ranges from 1 to 337 depending on the firm. We performed three different analyses. First we analyzed the service parity hypothesis by comparing a CLEC (#1008) with only 10 observations to the ILEC. We randomly selected 400 sample pairs from the permutation distribution and found the calculated Z to be 0.33858 and the corrected Z to be 0.70422. Next we analyzed the service parity hypothesis by comparing a CLEC (#1002) with 24 observations to the ILEC. Again, randomly selecting 400 sample pairs, we found the computed Z to be -0.5842 and the corrected Z to be -0.7023. Finally we compared a CLEC (#1025) with 131 observations to the ILEC. We randomly drew

400 sample pairs from the permutation distribution and found some unexpected results. The calculated Z was -1.619 but the corrected Z was -3.023. It is surprising to see this substantial of an adjustment for a sample as large as 131. Maybe no one really does know how large is large; or maybe 400 sample pairs in not nearly enough in view of the massive number of possible samples in this case. At any rate, the result is unanticipated; one would typically expect the LCUG-Z to be more accurate for larger sample sizes.

A final, perhaps discouraging, note: Each of these analyses took 15-20 minutes to run. While it is true that our program may not be the most efficient, it is also true that SAS is remarkably adroit at handling large sets of data. Thus we do not anticipate a significant improvement in running time unless we move from a personal computer (Pentium II) to a Sun or mainframe computer.

APPENDIX I

Programming Permutation Tests

Since there is a strong likelihood that someone at BANY is going to have to program these procedures, we offer the following summary of our programming logic.

1. Pool CLEC and ILEC samples.

2. Randomly draw one sample pair (two samples, one having n_{CLEC} and the other having the remaining n_{ILEC} observations) from the pooled data set.

3. Compute and store the LCUG-Z for this sample.

4. Repeat steps 3 and 4 for the remaining T - 1 sample pairs to be analyzed. (NOTE: in our program we put in a check at this step so as to prevent drawing more than one random sample composed of the same elements.)

5. Order the Zs computed and stored in step 4 from lowest to highest.

6. Compute the LCUG-Z for the observed sample and find its rank in the ordering determined in step 5.

7. Find the critical value of Z (Z_C) such that the probability that $Z < Z_C$ equals (rank from step 6)/T.

8. Repeat steps 2-7, say 10 times to get some indication of the robustness of Z_C.

9. Combine results: $P(Z < Z_C)$= (Σ ranks in each of the 10 runs)/10T

Code used by Colin Mallows

```
ppermLCUGt <- function(x, y, Nperm, .RS = .Random.seed)
{

# given two samples, computes the LCUG t, and gets Nperm-1 Monte Carlo # values of
this; hence the pperm  tail-area.

#

.Random.seed <<- .RS

xy <- c(x, y)
N <- length(xy)
m <- length(x)
n <- length(y)
tLCUG <- LCUGt(x, y)
permtLCUG <- 1:Nperm
permtLCUG[1] <- tLCUG
for(i in 2:Nperm) {
s <- sample(N, n)
yy <- xy[s]
xx <- xy[ - s]
permtLCUG[i] <- LCUGt(xx, yy)
      }

 (sum(permtLCUG >= tLCUG) - 0.5)/Nperm
}

LCUGt <- function(x, y)
{
m <- length(x)
n <- length(y)
(mean(y) - mean(x))/sqrt((var(x) * (m + n))/(m * n))
}
```

SAS Program used by John Jackson

We note that the consensus around here is that anyone trying to replicate this program in DBASE is not likely to be very successful. Thus BANY might wish to consider contracting out these calculations or updating its software.

```
/************************************************************/
/** Macro variables                              **/
/**  k       - k value                       **/
/**  m       - m value                        **/
/**  numsamp - number of samples                   **/
/**  numvars - number of variables                **/
/**  seed    - random number seed (0) uses clock value **/
/**  totobs  - m + k                        **/
/************************************************************/
  %LET K       = 11;
  %LET M       = 214;
  %LET NUMSAMP = 1000;
  %LET NUMVARS = 1;
  %LET SEED    = 0;
  %LET TOTOBS  = 225;

/************************************************************/
/** read in sample BANY data                          **/
/**LEC= 0 if ILEC, 1 if CLEC                           **/
/**                                                   **/
/************************************************************/
  DATA TEST;
 INPUT I LEC VAL1;
 DATALINES;
1    0    9.385
2    0    7.522
3    0    8.594
4    0    5.093
5    0    7.53
6    0    9.388
7    0    3.477
8    0    6.465
9    0    6.375
10    0    5.666
11    0    8.686
12    0    6.833
13    0    5.528
14    0    7.688
15    0    18.573
16    0    7.424
```

17	0	19.107
18	0	7.538
19	0	7.489
20	0	12.933
21	0	15.981
22	0	10.246
23	0	12.673
24	0	10.324
25	0	12.567
26	0	6.526
27	0	9.402
28	0	7.384
29	0	12.622
30	0	7.151
31	0	7.32
32	0	7.364
33	0	8.638
34	0	12.833
35	0	9.526
36	0	7.363
37	0	6.567
38	0	6.17
39	0	5.599
40	0	7.431
41	0	6.545
42	0	5.759
43	0	8.582
44	0	9.7
45	0	12.805
46	0	10.1
47	0	3.558
48	0	5.886
49	0	4.44
50	0	4.534
51	0	5.937
52	0	5.421
53	0	5.887
54	0	9.398
55	0	15.123
56	0	6.433
57	0	9.452
58	0	7.016
59	0	6.734
60	0	7.536
61	0	5.217
62	0	8.926

63	0	6.204
64	0	9.591
65	0	9.714
66	0	9.116
67	0	7.021
68	0	5.205
69	0	7.065
70	0	8.909
71	0	9.299
72	0	15.918
73	0	6.95
74	0	16.382
75	0	33.958
76	0	11.11
77	0	10.787
78	0	8.285
79	0	18.877
80	0	4.765
81	0	6.954
82	0	5.661
83	0	7.131
84	0	5.017
85	0	6.797
86	0	5.604
87	0	6.989
88	0	32.724
89	0	8.733
90	0	12.073
91	0	12.628
92	0	11.874
93	0	8.792
94	0	10.002
95	0	7.433
96	0	11.846
97	0	10.704
98	0	10.432
99	0	12.043
100	0	9.666
101	0	9.295
102	0	6.457
103	0	8.719
104	0	6.318
105	0	6.79
106	0	8.049
107	0	5.031
108	0	11.2

109	0	8.502
110	0	4.571
111	0	8.15
112	0	2.934
113	0	5.615
114	0	6.419
115	0	8.869
116	0	7.526
117	0	9.195
118	0	8.127
119	0	8.833
120	0	8.52
121	0	6.771
122	0	9.083
123	0	8.898
124	0	6.694
125	0	9.799
126	0	8.199
127	0	20.683
128	0	7.394
129	0	8.329
130	0	6.399
131	0	5.618
132	0	8.868
133	0	5.328
134	0	5.583
135	0	5.718
136	0	9.81
137	0	10.694
138	0	9.357
139	0	9.825
140	0	6.133
141	0	8.042
142	0	4.003
143	0	7.749
144	0	10.058
145	0	7.557
146	0	9.163
147	0	7.887
148	0	15.403
149	0	5.671
150	0	12.429
151	0	9.767
152	0	7.546
153	0	8.971
154	0	5.695

155	0	5.082
156	0	6.326
157	0	6.895
158	0	6.44
159	0	5.425
160	0	8.099
161	0	6.1
162	0	7.1
163	0	6.451
164	0	7.145
165	0	10.499
166	0	7.589
167	0	6.461
168	0	3.094
169	0	8.074
170	0	4.488
171	0	8.185
172	0	8.1
173	0	4.1
174	0	2.796
175	0	4.193
176	0	7.292
177	0	4.051
178	0	5.261
179	0	3.673
180	0	2.184
181	0	9.167
182	0	7.615
183	0	6.27
184	0	3.1
185	0	4.51
186	0	6.776
187	0	8.705
188	0	6.844
189	0	6.77
190	0	5
191	0	8.48
192	0	2.813
193	0	5.669
194	0	6
195	0	8
196	0	5.974
197	0	13.589
198	0	4.708
199	0	6.133
200	0	9.061

```
201   0   9.45
202   0   4.851
203   0   7.628
204   0   8.643
205   0   8.961
206   0   11.928
207   0   9.906
208   0   9.357
209   0   7.752
210   0   8.162
211   0   9.805
212   0   11
213   0   8
214   0   7.154
215   1   4.935
216   1   2.829
217   1   5.054
218   1   1.104
219   1   5.903
220   1   5.842
221   1   5.448
222   1   10.179
223   1   6
224   1   9.079
225   1   53.007

RUN;

/*********************************************************/
/** Find the Z for the data sample                     **/
/*********************************************************/
  PROC SUMMARY NWAY;
   VAR VAL1;
   CLASS LEC;
   OUTPUT OUT = SAMPSTAT
       MEAN = MEAN
       STD  = STD;
  RUN;

PROC PRINT;
  TITLE 'SAMPLE STATS';
RUN;

  DATA SAMPZ;
   SET SAMPSTAT;
   RETAIN MMEAN MSIGMA;
   IF LEC = 0 THEN
    DO;
     MMEAN = MEAN;
     MSIGMA = STD;
     DELETE;
    END;
   ELSE
    KMEAN = MEAN;
   Z = (KMEAN - MMEAN) / (MSIGMA * SQRT(1 / &K + 1 / &M));
   KEEP Z;
  RUN;

/**********************************************************/
/**  Generate &numvars X &numsamp arrays of k random     **/
/**  number uniformly distributed between 1 and &TOTOBS. **/
/**  For each of the &numvars sets, the &numsamp arrays  **/
/**  of k random numbers are unique.                     **/
/**********************************************************/
  DATA RANNUMS;
     ARRAY TEMP (&TOTOBS)  $1 _TEMPORARY_;
     ARRAY KSAVE (&NUMSAMP,&K) 3 _TEMPORARY_;
     ARRAY KVALS (&K) 3 _TEMPORARY_;
     ARRAY KOUT  (&K) ;
     /* loop for each variable */
       DO VAR = 1 TO &NUMVARS;

         /* initialize ksave */
         DO I = 1 TO &NUMSAMP;
             DO J = 1 TO &K;
                 KSAVE(I, J) = .;
             END;
         END;
     /* start of samples loop */


     NUMTOGO = &NUMSAMP;
         DO UNTIL (NUMTOGO = 0);
         /* gen k unique random numbers */
          K = &K;

           DO UNTIL (K=0);
           RANNUM = CEIL ( UNIFORM (&SEED) * &TOTOBS);
           /* check that num is unique */
           IF TEMP(RANNUM) = ' ' THEN
             DO;
                 TEMP(RANNUM) = '1';
```

```
            K = K - 1;
        END;
      END;

    /* load index of nonmissing in temp */
    /* as values of kvals             */
INDEX = 1;
    DO I = 1 TO &TOTOBS;
    IF TEMP(I) = '1' THEN
        DO;
            KVALS(INDEX) = I;
             TEMP(I) = ' ';
            INDEX + 1;
        END;
    END;

    /* check that sequence is unique */ MATCH = 'N';
 KROW = 1;
    DO WHILE (KSAVE(KROW, 1) ^= . & match ='N');
        DO COL = 1 TO &K UNTIL (MATCH = 'N');
            IF KVALS(COL) = KSAVE(KROW, COL) THEN
            MATCH = 'Y';
            ELSE
            MATCH = 'N';
        END;
        KROW + 1;
    END;
    IF MATCH = 'N' THEN    /* unique */
         DO;
            NUMTOGO + -1;
              DO COL = 1 TO &K;
               KSAVE(KROW, COL) = KVALS(COL);
              END;
         END;
    END;      /* end of samples loop - ksave loaded */
          /* output random numbers as */
          /* kout1 - kout&k          */

    DO SAMPNUM = 1 TO &NUMSAMP;
        DO J = 1 TO &K;
            KOUT(J) = KSAVE(SAMPNUM,J);
        END;
        OUTPUT;
    END;
END;              /* end of variables loop */
KEEP VAR SAMPNUM KOUT1-KOUT&K;
```

```
RUN;


/************************************************************/
/** Pull samples from data of size k for each of the   **/
/** &numvars variables based on the random numbers      **/
/** generated previously.  For each sample, calculate  **/
/** the kmean, ksigma, mmean, msigma and z             **/
/************************************************************/
  DATA ZVALS;
   ARRAY KOUT (&K) KOUT1-KOUT&K;
   ARRAY VAL  (&NUMVARS);
   ARRAY KVAL (&K);
   ARRAY MV (&M);
   SET RANNUMS;
   KINDEX = 1;
   MINDEX = 1;
   DO OBS = 1 TO &TOTOBS;
    PNT = OBS;
    SET TEST POINT = PNT;
    IF OBS = KOUT(KINDEX) THEN
     DO;
      KVAL(KINDEX) = VAL(VAR);
      KINDEX + 1;
      KINDEX = MIN(&K, KINDEX);
     END;
    ELSE
     DO;
      MV(MINDEX) = VAL(VAR);
      MINDEX + 1;
     END;
   END;
   KMEAN = MEAN(OF KVAL1 - KVAL&K);
   MMEAN = MEAN(OF MV1 - MV&M);
   KSIGMA = STD(OF KVAL1 - KVAL&K);
   MSIGMA = STD(OF MV1 - MV&M);
   Z = (KMEAN - MMEAN) / (MSIGMA * SQRT(1 / &K + 1 / &M));
   KEEP VAR SAMPNUM Z ;
  RUN;

/************************************************************/
/** Order the Z's                              **/
/************************************************************/
  PROC APPEND BASE = ZVALS
        DATA = SAMPZ;
  RUN;
```

```
PROC SORT DATA = ZVALS;
 BY Z;
RUN;

DATA PROB;
 SET ZVALS;
 IF SAMPNUM = . THEN
  DO;
   OBSNO = _N_;
   PROB = (_N_ - .5) / (&NUMSAMP + 1);
   zcorrect =probit(prob);
   OUTPUT;
  END;
 KEEP OBSNO Z PROB zcorrect;
RUN;

PROC PRINT;
TITLE ' Z-TEST OUTPUP ';
RUN;
```